

**Gilbert Deléage
Manolo Gouy
Alexandre de Brevern**

Bioinformatique

De la séquence à la structure des protéines

Cours et cas pratiques

3^e ÉDITION

DUNOD

© Dunod, 2013, 2015, 2021

11, rue Paul Bert, 92240 Malakoff
www.dunod.com

ISBN 978-2-10-082794-7

Le Code de la propriété intellectuelle n'autorisant, aux termes de l'article L. 122-5, 2° et 3° a), d'une part, que les « copies ou reproductions strictement réservées à l'usage privé du copiste et non destinées à une utilisation collective » et, d'autre part, que les analyses et les courtes citations dans un but d'exemple et d'illustration, « toute représentation ou reproduction intégrale ou partielle faite sans le consentement de l'auteur ou de ses ayants droit ou ayants cause est illicite » (art. L. 122-4).

Cette représentation ou reproduction, par quelque procédé que ce soit, constituerait donc une contrefaçon sanctionnée par les articles L. 335-2 et suivants du Code de la propriété intellectuelle.

Table des matières

Comment utiliser cet ouvrage	VII
Avant-propos	IX
1 La composition en acides aminés	1
1. Acides aminés et séquence	1
2. Informations déduites de la composition en acides aminés	4
2 Bases de données pour données de bases	7
1. Les banques de données généralistes	7
2. Une entrée SWISS-PROT	14
3. Les systèmes d'interrogation	18
3 La comparaison de deux séquences	21
1. Matrice de points	21
2. Matrice de substitution	26
4 Recherche dans les banques	33
1. Score de similitude entre séquences	33
2. Recherche globale ou locale	36
3. FASTA	37
4. BLAST	42
5 Alignement de séquences	47
1. Introduction	47
2. Comparaison de protéines homologues (algorithme global)	49
3. Meilleur chevauchement entre séquences (algorithme local)	52
4. Alignements multiples	55
5. Représentation « logo »	58

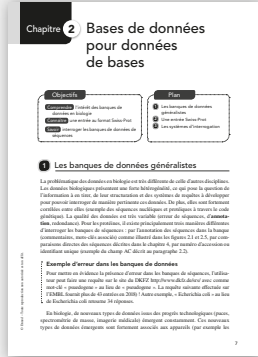
6	Bases théoriques de la phylogénie moléculaire	59
1.	Arbres phylogénétiques	59
2.	Arbre des espèces – arbres de gènes	63
3.	Modèle markovien de l'évolution moléculaire	66
4.	Choix des sites	72
5.	Matrices de taux de substitution entre séquences protéiques	74
6.	Distances évolutives entre paires de séquences	75
7	Algorithmes pour la phylogénie moléculaire	79
1.	Parcimonie	80
2.	Méthodes de distances	87
3.	Maximum de vraisemblance	91
4.	Estimation de la fiabilité d'un arbre par bootstrap	96
5.	Choix des méthodes de calcul d'arbres	99
8	Recherche de fonctions	101
1.	Définitions	101
2.	Détection de signatures de séquence (PROSITE)	102
3.	Recherche de fonction avec pondération par la fréquence	106
4.	Méthodes à base de profils	108
9	Profils physico-chimiques	113
1.	Pourquoi les profils physico-chimiques ?	113
2.	Hydrophobie-paramètres-construction du profil – interprétation	113
3.	Amphiphilie	116
4.	Accessibilité au solvant	117
10	Prédictions de structures secondaires	119
1.	Méthode « statistique empirique »	122
2.	Méthode information directionnelle (GOR)	125
3.	Méthode de recherche des plus proches voisins (NNM)	129
4.	Méthode auto-optimisée (SOPM)	132
5.	Méthode auto-optimisée avec alignements (SOPMA)	134

6. Méthodes d'apprentissage	134
7. Réseaux Neuronaux Artificiels	136
8. Machines à Vecteurs de Supports	140
9. Apprentissage profond	142
10. Mesures de qualité prédictive	148
11 Structures 3D	151
1. Principe des méthodes de détermination expérimentale	151
2. Le format PDB	153
3. Les différents modes de représentations	154
4. Classification de structures 3D	157
5. Comparaison de structures 3D	158
6. Énergétique moléculaire	160
7. Optimisation de structures 3D	163
12 Modélisation de structures 3D	165
1. Les méthodes d'enfilage des repliements (<i>threading</i>)	165
2. Modélisation par homologie	167
3. Les alphabets structuraux	176
4. Méthodes <i>ab initio</i>	188
13 Détection de sites 3D dans les protéines	191
1. Problématique	191
2. Méthode SuMO	191
Cas pratique d'analyse de séquences	195
Cas pratique de modélisation moléculaire de protéine par homologie	207
Conclusion	215
Bibliographie	217
Glossaire	223
Index	225

Comment utiliser cet ouvrage ?

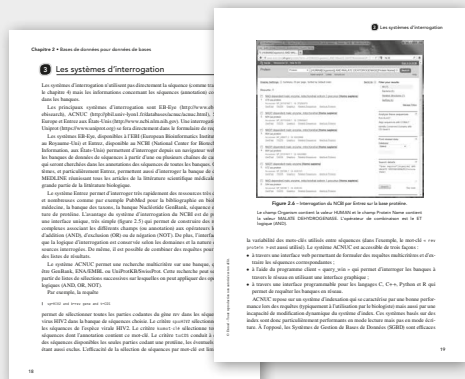
1 La page d'entrée de chapitre

Elle donne le plan du cours ainsi qu'un rappel des objectifs pédagogiques du chapitre.



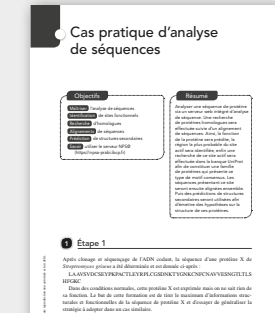
2 Le cours

Le cours, concis et structuré, expose le programme.



3 Cas pratique

Proposé en fin d'ouvrage, avec les réponses aux questions, pour appliquer les notions du cours.



4 Les rubriques



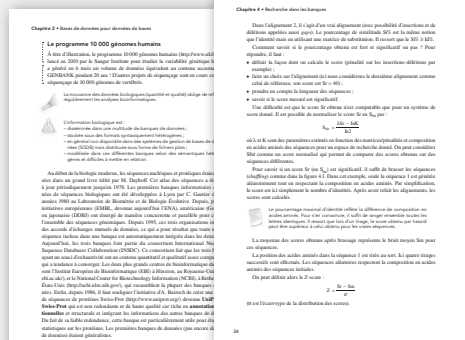
Une erreur
à éviter.



Une note,
un complément
d'information.



Un éclairage
sur une notion
(résumé, conseil, etc.)



5 La bibliographie

Elle regroupe les articles fondateurs de la discipline.

6 Le glossaire

Vous y trouverez les définitions des principales notions développées.

7 L'index

Outil indispensable pour trouver rapidement ce que l'on cherche.

Bibliographie

[illegible]

392

Glossaire

[illegible]

330

Index

[illegible]

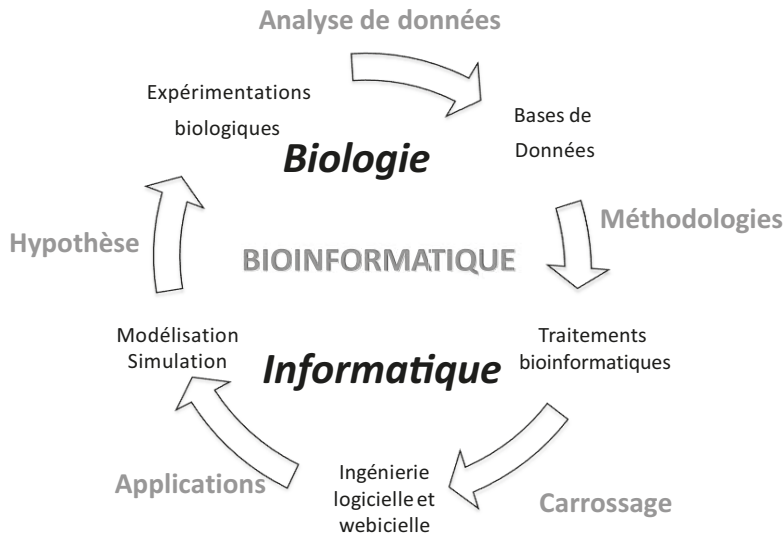
Avant-propos

La bioinformatique est une « interdiscipline » à la frontière de la biologie, de l'informatique et des mathématiques. Les systèmes biologiques sont très complexes et les techniques modernes d'investigation du monde biologique fournissent une vaste quantité de **données** expérimentales. Le but ultime de la bioinformatique est d'intégrer ces données d'origines très diverses pour modéliser les systèmes vivants afin de comprendre et prédire leurs comportements (biologie systémique ou biologie des systèmes) dans des conditions de fonctionnement normales ou pathologiques. Ainsi, à titre d'exemple, le séquençage à très haut débit offre la possibilité de connaître de manière personnalisée le **génome** de chacun. Pour tirer le bénéfice de cette connaissance, il faut développer et appliquer de nouvelles méthodes d'analyse bioinformatique qui permettent d'extraire l'information utile cachée dans la séquence du génome et, de manière plus générale, des données biologiques à grande échelle issues des progrès de l'expérimentation et des technologies de l'automatique. La bioinformatique est donc étroitement couplée à ses applications. Bon nombre de bioinformaticiens ne travaillent pas dans des laboratoires formellement estampillés « bioinformatique ». La bioinformatique et la modélisation procèdent selon un cercle vertueux (schématisé page suivante) dans lequel le point de départ est l'expérimentation biologique (un séquençage par exemple), les données produites sont ensuite organisées dans des dépôts de données (banques ou bases de données). Les méthodes d'analyse qui utilisent ces données sont développées par les bioinformaticiens souvent en association avec des informaticiens et mathématiciens. Pour que ces méthodes permettent le traitement ultérieur des données, il est nécessaire de « carrosser » ces méthodes (sous forme de logiciels ou serveurs web) afin de permettre au biologiste de les utiliser pour émettre de nouvelles hypothèses qui seront testées et qui généreront de nouvelles données.

Aujourd'hui tout projet de biologie comporte une étape d'analyse bioinformatique des données. Par conséquent, un biologiste passe environ 20-30 % de son temps à utiliser des outils bioinformatiques. D'ailleurs, il est remarquable de voir que les deux articles les plus cités en biologie sont ceux décrivant BLAST et CLUSTAL, deux méthodes bioinformatiques largement utilisées par les biologistes.

Ce livre décrit de manière simple les tâches courantes de la bioinformatique qu'un biologiste/biochimiste doit savoir traiter par lui-même sans avoir recours au spécialiste afin de répondre à des questions usuelles comme :

- Comment extraire des informations pertinentes dans les banques de données biologiques ?
- Est-ce qu'une nouvelle séquence a déjà été complètement ou partiellement répertoriée ?
- Est-ce que ce gène appartient à une famille connue ?

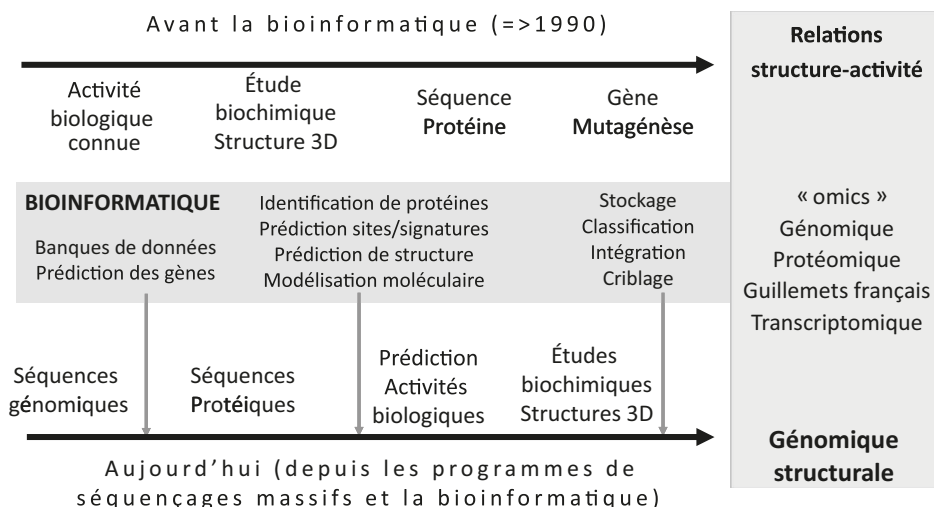


- Existe-t-il d'autres gènes homologues ?
- Est-ce que deux séquences correspondent à deux gènes homologues ?
- Existe-t-il des résidus essentiels à la fonction ?
- Alignement multiple, quel outil ? Pour quoi faire ? Établissement de consensus.
- Quelle peut être la fonction d'une protéine (prédit d'après sa séquence, sa structure...) ?
- Recherche de sous-motifs communs à un ensemble de séquences.
- Recherche de régions contenant des séquences répétées.
- Recherche d'hélices ou de brins dans les protéines.
- Comment construire un modèle tridimensionnel de protéine ?
- Optimisation et comparaison de structures 3D.
- Quelle est la charge globale d'une protéine à un pH donné ?

Ce livre n'a pas la prétention d'être exhaustif (il se limite d'une manière générale aux protéines, mais les **algorithmes** sont souvent très proches de ceux développés pour les acides nucléiques). Il a été rédigé afin de faciliter la compréhension des approches, méthodes, algorithmes et **implémentations** les plus courantes en bioinformatique moléculaire et structurale. À ce titre, il est parfois simplificateur et doit être considéré comme une introduction à la bioinformatique moléculaire et structurale. Il s'adresse donc aux étudiants de biologie/biochimie, de niveau licence, master ou classes préparatoires, ou bien aux biologistes qui souhaitent s'initier et comprendre les méthodes sous-jacentes aux programmes afin d'estimer la qualité de leurs analyses.

La logique suivie dans le livre est de partir des séquences de protéines pour aller vers leurs structures secondaires, leurs structures tridimensionnelles et finir par leurs fonctions. Elle suit la stratégie actuelle d'analyse d'une question biologique qui a été revisitée du fait de l'avènement de la bioinformatique et des séquençages massifs. La

bioinformatique moléculaire a pour première mission de « faire parler cette séquence » pour en tirer le maximum d'informations selon le schéma suivant :



La plupart des images des structures 3D présentées ont été générées à l'aide du logiciel AnTheProt pour Windows (<http://antheprot-prabi.ibcp.fr>).

En fin d'ouvrage, un exercice de mise en pratique de l'analyse de séquence est fourni avec son corrigé. De même, un exercice avec corrigé concerne la modélisation moléculaire par homologie. Un quizz en ligne est disponible à l'adresse suivante : https://www.gdeleage.fr/prof/bio_info.php?choix=BIOINFORMATIQUE.

Les auteurs remercient Christophe Combet et Céline Brochier pour la relecture du livre.